Automating Marine Mammal Detection in Aerial Images Captured During Wildlife Surveys: a Deep Learning Approach

Frederic Maire[1], Luis Mejias Alvarez[1], and Amanda Hodgson[2]

[1] Science and Engineering Faculty,
Queensland University of Technology,
Brisbane, Australia
f.maire,luis.mejias@qut.edu.au
[2] Murdoch University Cetacean Research Unit,
Murdoch University,
Perth, Australia
a.hodgson@murdoch.edu.au

Abstract. Aerial surveys conducted using manned or unmanned aircraft with customized camera payloads can generate a large number of images. Manual review of these images to extract data is prohibitive in terms of time and financial resources, thus providing strong incentive to automate this process using computer vision systems. There are potential applications for these automated systems in areas such as surveillance and monitoring, precision agriculture, law enforcement, asset inspection, and wildlife assessment. In this paper, we present an efficient machine learning system for automating the detection of marine species in aerial imagery. The effectiveness of our approach can be credited to the combination of a well-suited region proposal method and the use of Deep Convolutional Neural Networks (DCNNs). In comparison to previous algorithms designed for the same purpose, we have been able to dramatically improve recall to more than 80% and improve precision to 27% by using DCNNs as the core approach.

1 Introduction

Aerial surveys conducted in light aircraft are a common technique for monitoring various species of wildlife throughout the world (e.g. [6] [21] [25]). As imaging technologies improve, researchers are moving towards replacing human observers with camera systems (e.g. [17] [8] [36]), and replacing manned aircraft with Unmanned Aerial Vehicles (UAVs) ([35] [16]). These technologies have the potential to improve safety, reduce costs, increase the reliability of the data, and allow surveys to be conducted in remote or inaccessible areas.

One of the main limitations in adopting these new methods is the onerous task of reviewing the captured images to extract data on wildlife sightings. Hodgson et al. [16], for example, conducted a trial to assess the potential for using UAVs to survey dugongs (Dugong dugon) in their marine habitat. They captured

overlapping images (in the visual spectrum) continuously along the survey route and manually reviewed the images post survey to determine the distribution and abundance of dugongs. Surveys using these methods result in tens of thousands of images that take months to review.

Although there are many researchers facing the challenge of reviewing large image datasets from aerial surveys, there is no software readily available that automates this process. The main challenge of detecting animals in aerial images is the environmental conditions like the turbidity of the water and the presence of wave-crests, and their effects on illumination within the images. Despite these difficulties, computer vision is an attractive solution given it offers a rich and permanent source of information, and it is easily generalizable to many types of EO sensors and aircraft.

The few published works in this area include those by Advanced Coherent Technologies who have developed techniques to detect and track whales using multispectral cameras and computer vision [24, 23, 27, 28]. However the algorithms used in this system are not available to other researchers. Groom et al. [13] describe an object based image analysis method for surveying marine birds that reduces the number of images requiring manual review.

The algorithms presented in [20] and [19] were designed to automate the detection of dugongs in images using ad-hoc features based on color segmentation and blob shape analysis. While still under development, these two approaches highlighted the potential benefits of pattern recognition to this domain. However, these algorithms had moderate recall and comparatively low precision, and depended on the user to tune critical parameters.

An alternative to handcrafting image features is to learn them from annotated image datasets. In this paper, we introduce an approach aimed at automating the review process (or parts of it) based on Deep Learning, leveraging off the enormous momentum created by the deep learning community. We have been able to dramatically improve recall (achieving more than 80%) and improve precision (reaching 27%) by using Deep Convolutional Neural Networks (DCNNs) as a core approach. Here we compare two different deep learning architectures and advocate a simple *region proposals* method. This method involves performing clustering in the 5D space that consists of color information and image location.

This paper is structured as follows. Section 2 outlines the different components in the pipeline of our detector. Section 3 explains the *region proposals* module. Section 4 gives details about the DCNNs investigated. Section 5 describes the experiments carried out. The paper concludes with a discussion on the applicability of our system to other scenarios.

2 System Overview

Figure 1 illustrates the high level view of our dugong detector. We combine the efficiency of Simple Linear Iterative Clustering (SLIC), a region proposals algorithm reviewed in Section 3, and the generosity of deep convolutional neural networks in a pipeline to detect dugongs in large images.

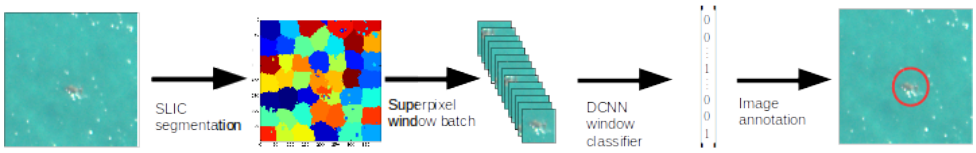

Fig. 1. The actual size of the input images is 6034×4012 pixels. For clarity, the input image used in the diagram is only 400×400. A dugong can be seen close to the center. The pipeline of the classifier starts with segmentation of the input image in superpixels using the SLIC algorithm. In the next step, a batch of windows centred at the superpixels is generated. This batch of windows is then fed to a deep convolutional neural network. Finally, a copy of the input image is annotated with the position of the detected dugongs.

To build this detector, we extracted training examples from an annotated image collection in a format suitable for a DCNN as sketched in Figure 2. The initial database of positive examples was enlarged by applying rotations and scaling transforms to this dataset (see Figure 3). We then sampled windows away from the marked locations to create an initial set of negative examples. Better negative examples (more ambiguous windows) were also obtained by training a first generation DCNN, and evaluating this DCNN on the training images. The false positives produced from this process replaced some of the initial negative window examples. A second generation DCNN was then trained on this updated database.

3 Region proposals

Since the early work on face detection, multi-scale scanning has been a popular approach for object detection in natural scenes [14]. A sliding window of fixed dimensions sweeps across the image, and at each offset is fed to a classifier to detect the presence of objects of a specific category such as faces [26, 32], pedestrians [29, 22] and vehicles [30]. To find objects of different sizes, the image is sub-sampled in a pyramid while the sliding window keeps the same dimensions. CNNs have been used as the classifier of such detection systems for at least two decades.

From a computational complexity point of view, the size of the neural network is a predominant factor determining the running time as the evaluation of the window classifier is the major time cost. Unfortunately, CNNs do not have an amortized cost like the Viola-Jones detector [34]. For a 20 Megapixel image, millions of candidate windows have to be processed. Even for CNNs that only have two convolutional and pooling layers, this cost becomes an issue [18].

Nevertheless, experimental results have demonstrated that deep convolutional neural networks (DCNN) are superior to shallow models for complex tasks [4]. An alternative to building a sliding-window detector is to operate a pipeline chaining a *region proposals* module and a CNN [14].

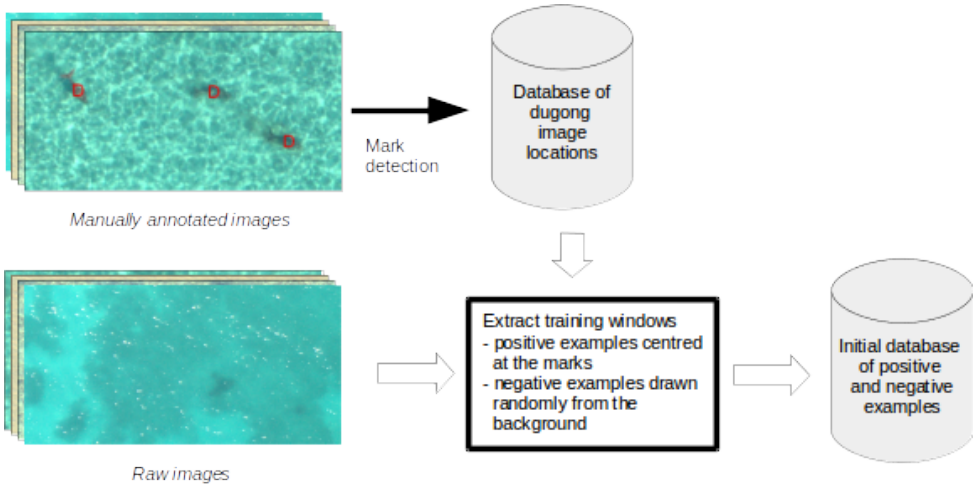

Fig. 2. The image collection contains 251 images in ppm format. The file size of each image is 72.6MB. Each image was annotated by an expert, then saved in a compressed jpg file. The input of the DCNN classifier is a 64×64 color window. An initial database of labeled windows was obtained by extracting 100×100 windows centred on the marks made by the experts for the positive examples, and windows of the same size randomly picked away from the marks for the negative examples.

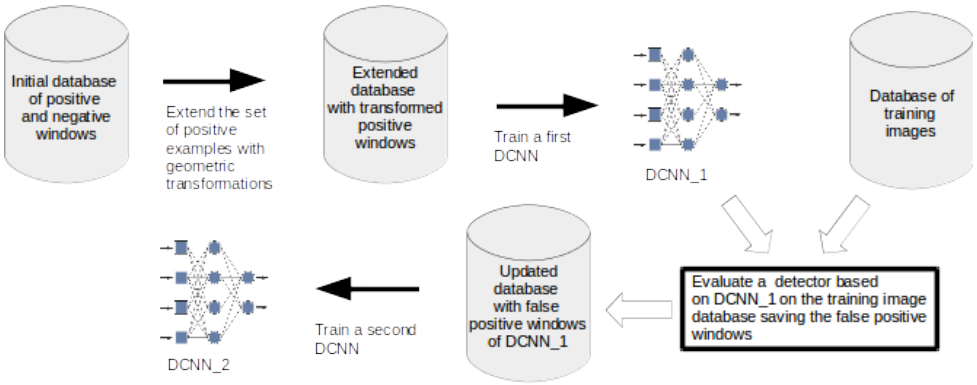

Fig. 3. Two generations of DCNNs are trained. The role of the first DCNN is to identified negative windows that are hard to classify. These difficult examples enrich the initial database, and allow to re-train DCNNs with better performance.

A number of recently introduced methods generate category independent region proposals. These generic methods include objectness [2], selective search [31], category-independent object proposals [9], constrained parametric min-cuts (CPMC) [7], multi-scale combinatorial grouping [3], and low-level image segmentation methods like [10], [33] and [1].

Felzenszwalb's graph based method [10] is a fast 2D segmentation algorithm that has only a single scale parameter to tune. We found that in practice the actual size and number of segments was quite sensitive to the local contrast. Quickshift [33] is another recent 2D image segmentation algorithm, based on kernelized mean-shift, that seeks local mode in the the 5D space consisting of color information and image location. Quickshift computes a hierarchical segmentation on multiple scales simultaneously and requires tuning two main parameters. SLIC [1] also operates in the same 5D space and performs clustering in this space using K-means. The clustering method is simple and very efficient. Moreover, it is particularly well suited to our application. Indeed, as we know the expected apparent size of the objects of interest (in our case, dugongs), we can select the compactness parameter (that trades off color-similarity and proximity), and the number of centers for K-means so that the superpixels are approximately the size of a dugong.

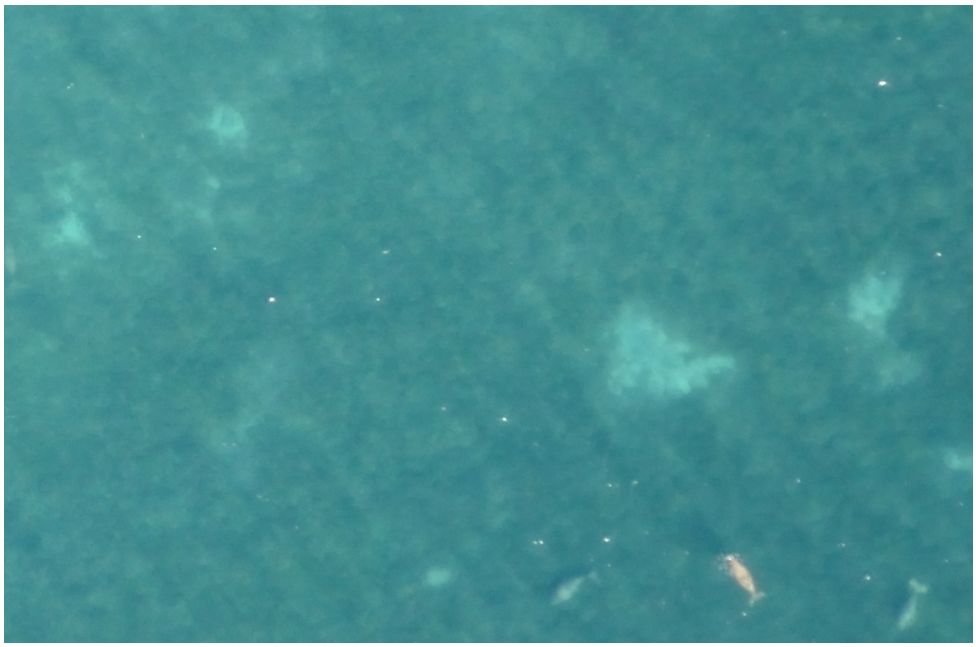

Fig. 4. In this 1860 × 1224 sub-image of a larger aerial image, we can distinguish 3 dugongs on the bottom right. One of the dugongs is surfacing and has a strong colour contrast with the background. The other two dugongs, on the left and on the right of the surfacing dugong, have a clear outline, but their colour is less distinctive.

The (approximate) number of superpixels in the segmented output image is a parameter n of the SLIC algorithm that we set to the number of dugongs that could be packed on a regular grid (imagine a dugong dormitory). That is, we set

$$n = \left(\frac{\sqrt{w \times h}}{d} \right)^2$$

where d is the expected side length of a bounding box for a dugong, and w and h are respectively the width and height of the full image.

Figure 4 shows part of an aerial photo taken under favourable weather conditions. At the bottom right of the image, we can see three dugongs grazing. In Figure 5, each coloured blob corresponds to a superpixel of the SLIC segmentation. The silhouettes of the dugongs are clearly visible. Their contours can be see in Figure 6.

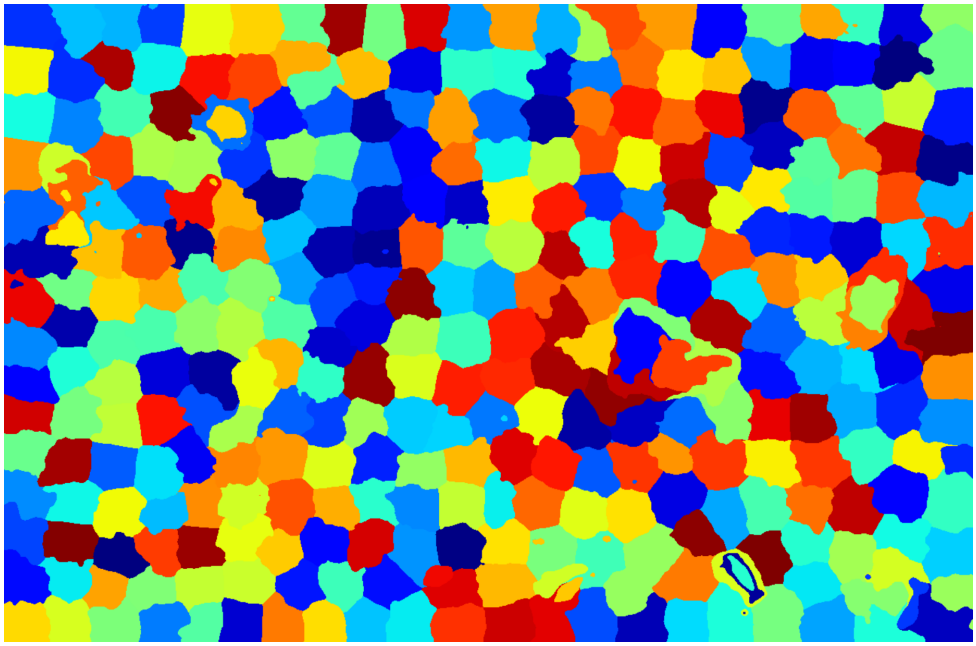

Fig. 5. SLIC segmentation of the image displayed in Figure 4. The bodies of the three dugongs correspond to three superpixels.

4 Convolutional Neural Network Architecture

As we had a larger dataset than the one described in [18], we were able to train more complex CNNs. In [18], the proposed CNN starts with two LeNet convolutional layers with hyperbolic transfer functions followed by a hidden layer, and

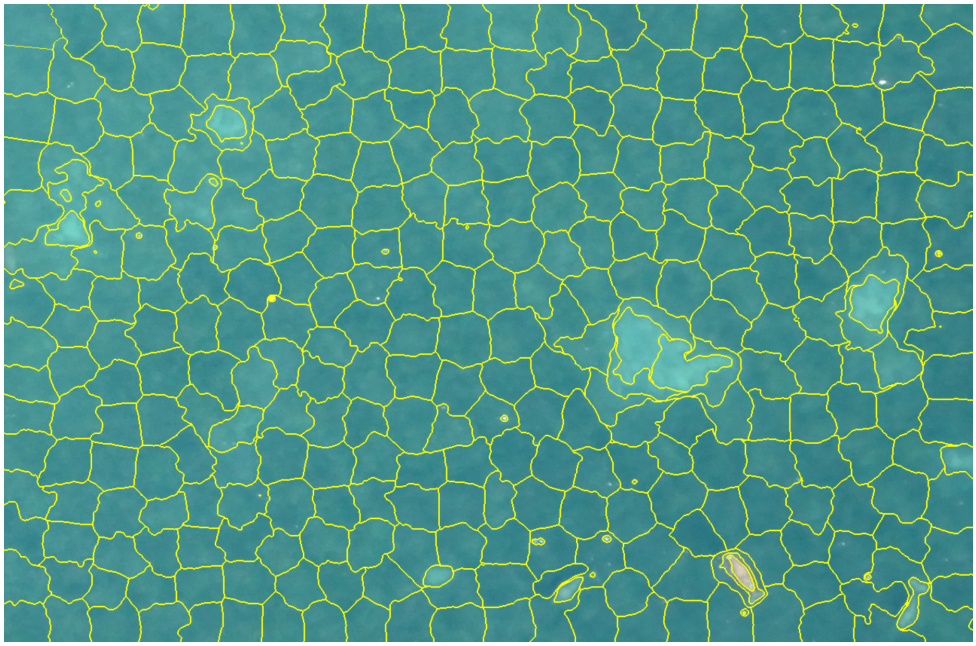

Fig. 6. Contour image of the SLIC segmentation of the image shown in Figure 4. We can observe that the contours of the bodies of the dugongs match superpixels.

finally a logistic regression layer. The output from the latter is a prediction of whether or not the input window contains a dugong. Our CNNs have three convolutional layers and use either rectilinear or maxout transfer functions. Adding a convolutional layer reduced the classification error by 3 percents.

We leveraged the Pylearn2 framework [11] to implement our deep CNNs. However, a few low-level functions had to be written using the Theano library [5] on which Pylearn2 is built.

We compared two CNN architectures for our application. The first architecture replaces the hyperbolic transfer function with a rectifier activation function $x \mapsto \max(0, x)$. This choice is mainly motivated by the more efficient computation (it used only elementary comparisons) associated with the rectified linear function compared to the hyperbolic transfer function. The use of the rectilinear function also eliminates the vanishing gradient problem associated with sigmoidal transfer functions.

The second architecture we experimented with was a deep convolutional neural network with a Maxout activation function [12]. In a convolutional network, a Maxout feature map takes the maximum across a finite number of affine feature maps (which are identical to those found in the convolutional layers of the first architecture). As the intersection of a number of half-spaces is a convex region, a single Maxout unit can be interpreted as making a piecewise linear approximation to an arbitrary convex function like the rectified linear function,

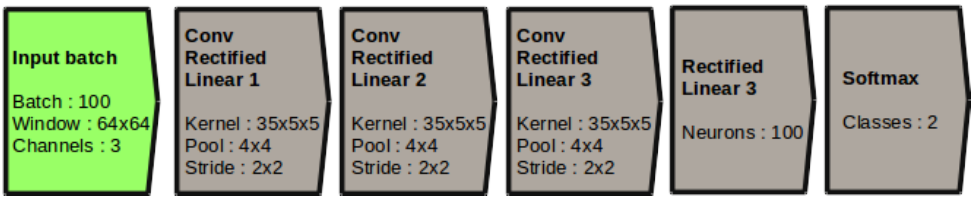

Fig. 7. Our first type of convolutional neural network starts with three convolutional layers followed by a hidden layer, and finally a logistic regression layer. The output from the latter is a prediction of whether or not the input window contains a dugong.

the absolute value function or a quadratic function. In a sense, Maxout networks learn their activation functions.

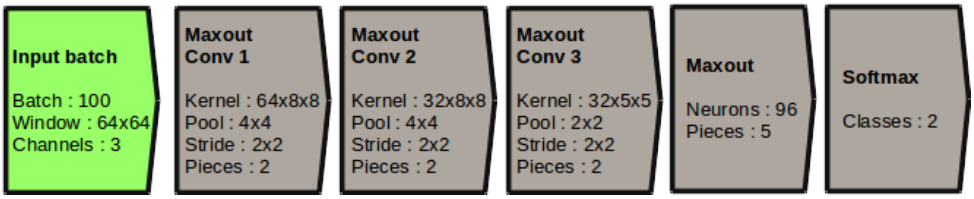

Fig. 8. The maxout convolutional neural network starts with three maxout convolutional layers followed by a maxout hidden layer, and finally the same logistic regression layer as in the convonet model of Figure 7.

Maxout networks work best when trained with Dropout [15]. Dropout emulates an inexpensive form of bagging by training a large ensemble of models that share parameters and approximately averaging these models' predictions. Dropout achieves this by randomly dropping units from the neural network during training (reducing the number of active connections). During training, Dropout samples from an exponential number of different reduced networks. During exploitation of the trained network, the effect of averaging the predictions of all these reduced networks is approximated by using the whole network with scaled down weights [12].

5 Experimental Results

The DCNNs were trained with weight decay using Stochastic Gradient Descent with a batch size of 100. The size of the layers of Rectilinear CNNs was determined empirically. The learning rate was 5.0×10^{-4}, and the final momentum was 0.8. The Maxout CNNs were trained with a learning rate of 1.0×10^{-3}. The weight include probability was set to 0.5. The size of the layers of the Maxout CNNs was constrained by the GPU implementation in Pylearn2.

We evaluated the detectors by computing the *precision*, *recall* and *F1-score* on a test set of large images. We use *TP*, *FP* and *FN* to denote respectively the number of true positives, false positives and false negatives respectively. By definition, the precision is *TP / (TP + FP)*, the recall is *TP / (TP + FN)*, and the *F1-score* is the harmonic mean of the precision and the recall. That is, *2 TP / (2 TP + FP + FN)*.

Table 1. Best detection results for the two types of DCNN

DCNN	TP	FP	FN	Precision	Recall	F1-score
First Generation Maxout	45	1245	6	0.0348	0.8823	0.0671
Second Generation Maxout	41	230	10	0.1512	0.8039	0.2546
First Generation Rectilinear	45	1909	6	0.0230	0.88235	0.04488
Second Generation Rectilinear	41	110	10	0.27152	0.8039	0.4059

Table 1 shows the benefit of training a second generation network. Although the recall performance of the second generation CNN decreased slightly compared to the first generation network, the precision significantly improved. The F1-score increased dramatically for the two types of networks considered. Figure 9 shows that the training of the second generation CNN takes more time but converges to the same error level as the first generation CNN.

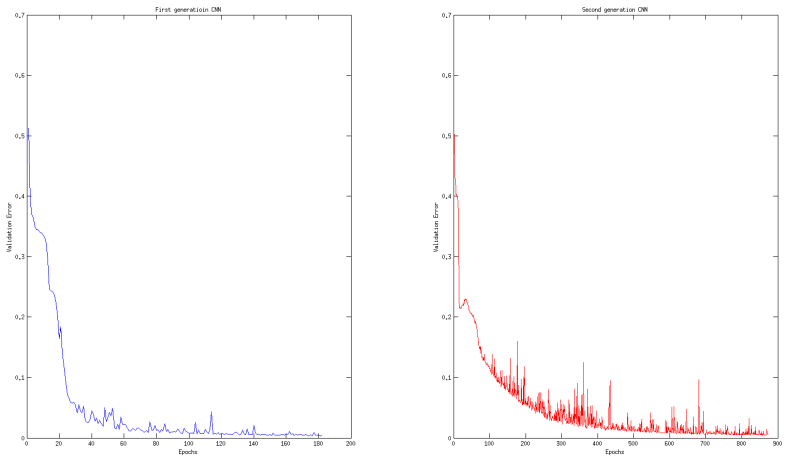

Fig. 9. Validation error plot of the two generations of CNNs. Starting at around 50% the validation error decreases progressively to less than 0.5%. The training of the second generation DCNN requires more epochs, probably because its training set is harder.

Training Maxout networks on a GPU takes only half a day, but getting the Pylearn2 software to run with a GPU can be tricky. Training of the Rectilinear DCNN takes about a week on a desktop PC (3GHz, 16GB of RAM).

In order to allow other researchers to make an objective and meaningful comparison between different approaches to the dugong detection challenge, we are releasing a dataset in HDF5 format and Python code to create Numpy arrays from this dataset, as well as Python scripts to replicate our experiments. The dataset and some scripts can be accessed at `https://cloudstor.aarnet.edu.au/plus/public.php?service=files&t=599e31295e83e074ac9e11d9d7a96922`

6 Conclusion

In this paper, we have demonstrated that the combination of region proposals based on SLIC segmentation and classification based on CNNs is well suited for the processing of aerial images from marine fauna surveys. In situations where photos are taken at a known height, and the altitude of the aircraft stays relatively constant, there is little variation in the apparent size of the animals, and no pyramidal analysis of the images is required. Our experimental results showed that deep architectures (three convolutional layers) do help improve detection performance.

As more image data is collected, the training set will grow, and better detection performance can be expected by retraining DCNNs. Marine biologists will no longer need to review whole images (such as the 20 Mega pixel images used in this project), but will instead check the windows labeled positively by the DCNN detector. This process should dramatically reduce the time required to extract data from images captured during aerial surveys. The corrected labeled windows can then be added to the training set database to continually improve the performance of the detector.

The dataset that we are releasing will allow the scientific community to objectively assess the research progress in dugong detection methods and potentially apply these methods to other image analysis problems.

References

1. Radhakrishna Achanta, Appu Shaji, Kevin Smith, Aurelien Lucchi, Pascal Fua, and Sabine Susstrunk. Slic superpixels compared to state-of-the-art superpixel methods. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 34(11):2274–2282, 2012.
2. Bogdan Alexe, Thomas Deselaers, and Vittorio Ferrari. Measuring the objectness of image windows. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 34(11):2189–2202, 2012.
3. Pablo Arbelaez, Jordi Pont-Tuset, Jon Barron, Ferran Marques, and Jitendra Malik. Multiscale combinatorial grouping. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 328–335. IEEE, 2014.

4. Yoshua Bengio, Aaron C. Courville, and Pascal Vincent. Unsupervised feature learning and deep learning: A review and new perspectives. *CoRR*, abs/1206.5538, 2012.
5. James Bergstra, Olivier Breuleux, Frédéric Bastien, Pascal Lamblin, Razvan Pascanu, Guillaume Desjardins, Joseph Turian, David Warde-Farley, and Yoshua Bengio. Theano: a cpu and gpu math expression compiler. *Proceedings of the Python for Scientific Computing Conference (SciPy)*, 2010.
6. Peter M. Calverley and Colleen T. Downs. Habitat use by nile crocodiles in ndumo game reserve, south africa: a naturally patchy environment. *Herpetologica*, 70(4):426–438, 2014.
7. Joao Carreira and Cristian Sminchisescu. Constrained parametric min-cuts for automatic object segmentation. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 3241–3248. IEEE, 2010.
8. Paul B. Conn, Jay M. Ver Hoef, Brett T. McClintock, Erin E. Moreland, Josh M. London, Michael F. Cameron, Shawn P. Dahle, and Peter L. Boveng. Estimating multispecies abundance using automated detection systems: ice-associated seals in the bering sea. *Methods in Ecology and Evolution*, 5(12, Sp. Iss. SI):1280–1293, 2014.
9. Ian Endres and Derek Hoiem. Category independent object proposals. In *Computer Vision–ECCV 2010*, pages 575–588. Springer, 2010.
10. Pedro F. Felzenszwalb and Daniel P. Huttenlocher. Efficient graph-based image segmentation. *International Journal of Computer Vision*, 59(2):167–181, 2004.
11. Ian J. Goodfellow, David Warde-Farley, Pascal Lamblin, Vincent Dumoulin, Mehdi Mirza, Razvan Pascanu, James Bergstra, Frédéric Bastien, and Yoshua Bengio. Pylearn2: a machine learning research library. *ArXiv e-prints*, August 2013.
12. Ian J. Goodfellow, David Warde-Farley, Mehdi Mirza, Aaron Courville, and Yoshua Bengio. Maxout networks. *arXiv preprint arXiv:1302.4389*, 2013.
13. Geoff Groom, Michael Stjernholm, Rasmus Due Nielsen, Andrew Fleetwood, and Ib Krag Petersen. Remote sensing image data and automated analysis to describe marine bird distributions and abundances. *Ecological Informatics*, 14(0):2–8, 2013.
14. Chunhui Gu, Joseph J. Lim, Pablo Arbeláez, and Jitendra Malik. Recognition using regions. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 1030–1037, June 2009.
15. Geoffrey E. Hinton, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan R Salakhutdinov. Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580*, 2012.
16. Amanda Hodgson, Natalie Kelly, and David Peel. Unmanned aerial vehicles (uavs) for surveying marine fauna: a dugong case study. *PLoS ONE*, 8(11):e79556, 2013.
17. William R. Koski, Tannis A. Thomas, Dale W. Funk, and A. Michael Macrander. Marine mammal sightings by analysts of digital imagery versus aerial surveyors: a preliminary comparison. *Journal of Unmanned Vehicle Systems*, 01(01):25–40, 2013.
18. Frederic Maire, Luis Mejias, and Amanda Hodgson. A convolutional neural network for automatic analysis of aerial imagery. In Lei Wang Wang, Philip Ogunbona, and Wanqing Li, editors, *Digital Image Computing: Techniques and Applications (DICTA 2014)*, Wollongong, New South Wales, Australia, 2014.
19. Frederic Maire, Luis Mejias, Amanda Hodgson, and Gwenael Duclos. Detection of dugongs from unmanned aerial vehicles. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2013)*, Tokyo Big Sight, Tokyo, 2013.

20. Luis Mejias, Gwenael Duclos, Amanda Hodgson, and Frederic Maire. Automated marine mammal detection from aerial imagery. In *OCEANS '13 MTS/IEEE*, Town and Country Resort Hotel, San Diego, CA, 2013.
21. Jean-Simon Michaud, Nicholas C. Coops, Margaret E. Andrew, Michael A. Wulder, Glen S. Brown, and Gregory J. M. Rickbeil. Estimating moose (alces alces) occurrence and abundance from remotely derived environmental indicators. *Remote Sensing of Environment*, 152:190–201, 2014.
22. Wanli Ouyang and Xiaogang Wang. Joint deep learning for pedestrian detection. In *Computer Vision (ICCV), 2013 IEEE International Conference on*, pages 2056–2063. IEEE, 2013.
23. Yuliya Podobna, Jon Schoonmaker, Cynthia Boucher, and Daniel Oakley. Optical detection of marine mammals. In *Proc. SPIE 7317, Ocean Sensing and Monitoring*, volume 7317, 2009.
24. Yuliya Podobna, James Sofianos, Jon Schoonmaker, Dustin Medeiros, Cynthia Boucher, Daniel Oakley, and Steve Saggese. Airborne multispectral detecting system for marine mammals survey. In *Proc. SPIE 7678, Ocean Sensing and Monitoring II, 76780G.*, April 20, 2010 2010.
25. Silje L. Rekdal, Rikke Guldborg Hansen, David Borchers, Lutz Bachmann, Kristin L. Laidre, Oystein Wiig, Nynne Hjort Nielsen, Sabrina Fossette, Outi Tervo, and Mads Peter Heide-Jorgensen. Trends in bowhead whales in west greenland: Aerial surveys vs. genetic capture-recapture analyses. *Marine Mammal Science*, 31(1):133–154, 2015.
26. Henry A Rowley, Shumeet Baluja, and Takeo Kanade. Neural network-based face detection. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 20(1):23–38, 1998.
27. Jon Schoonmaker, Yuliya Podobna, Cynthia Boucher, James Sofianos, Daniel Oakley, Dustin Medeiros, and Jose Lopez. The utility of automated electro-optical systems for measuring marine mammal densities. In *OCEANS 2010*, pages 1–6, Sept.
28. Jon Schoonmaker, Tami Wells, Gary Gilbert, Yuliya Podobna, Irina Petrosyuk, and Joseph Dirbas. Spectral detection and monitoring of marine mammals. In *SPIE 6946, Airborne Intelligence, Surveillance, Reconnaissance (ISR) Systems and Applications V, 694606*, 2008.
29. Pierre Sermanet, Koray Kavukcuoglu, Soumith Chintala, and Yann LeCun. Pedestrian detection with unsupervised multi-stage feature learning. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 3626–3633. IEEE, 2013.
30. Luo-Wei Tsai, Jun-Wei Hsieh, and Kuo-Chin Fan. Vehicle detection using normalized color and edge map. *Image Processing, IEEE Transactions on*, 16(3):850–864, 2007.
31. Jasper RR Uijlings, Koen EA van de Sande, Theo Gevers, and Arnold WM Smeulders. Selective search for object recognition. *International journal of computer vision*, 104(2):154–171, 2013.
32. Régis Vaillant, Christophe Monrocq, and Yann Le Cun. Original approach for the localisation of objects in images. *IEE Proceedings-Vision, Image and Signal Processing*, 141(4):245–250, 1994.
33. Andrea Vedaldi and Stefano Soatto. Quick shift and kernel methods for mode seeking. In *Computer Vision–ECCV 2008*, pages 705–718. Springer, 2008.
34. Paul Viola and Michael Jones. Rapid object detection using a boosted cascade of simple features. In *Computer Vision and Pattern Recognition, 2001. CVPR 2001.*

Proceedings of the 2001 IEEE Computer Society Conference on, volume 1, pages I–511. IEEE, 2001.

35. Adam C. Watts, John H. Perry, Scot E. Smith, Matthew A. Burgess, Benjamin E. Wilkinson, Zoltan Szantoi, Peter G. Ifju, and H. Franklin Percival. Small unmanned aircraft systems for low-altitude aerial surveys. *Journal of Wildlife Management*, 74(7):1614–1619, 2010.
36. Scott Wilson, Ron Bazin, Wendy Calvert, Terry Doyle, Stephen D. Earsom, Stephen A. Oswald, and Jennifer M. Arnold. Abundance and trends of colonial waterbirds on the large lakes of southern manitoba. *Waterbirds*, 37(3):233–244, 2014.